Legal Deposit Web Archives and the Digital Humanities: A Universe of Lost Opportunity?

Paul Gooding
p.gooding@uea.ac.uk
University of East Anglia, United Kingdom

Melissa Terras
m.terras@ed.ac.uk
University of Edinburgh, United Kingdom

Linda Berube
l.berube@uea.ac.uk
University of East Anglia, United Kingdom

Introduction

Legal deposit libraries have archived the web for over a decade. Several nations, supported by legal deposit regulations, have introduced comprehensive national domain web crawling, an essential part of the national library remit to collect, preserve and make accessible a nation's intellectual and cultural heritage (Brazier, 2016). Scholars have traditionally been the chief beneficiaries of legal deposit collections: in the case of web archives, the potential for research extends to contemporary materials, and to Digital Humanities text and data mining approaches. To date, however, little work has evaluated whether legal deposit regulations support computational approaches to research using national web archive data (Brügger, 2012; Hockx-Yu, 2014; Black, 2016).

This paper examines the impact of electronic legal deposit (ELD) in the United Kingdom, particularly how the 2013 regulations influence innovative scholarship using the Legal Deposit UK Web Archive. As the first major case

study to analyse the implementation of ELD, it will address the following key research questions:

- Is legal deposit, a concept defined and refined for print materials, the most suitable vehicle for supporting DH research using web archives?
- How does the current framing of ELD affect digital innovation in the UK library sector?
- How does the current information ecology, including not for-profit archives, influence the relationship between DH researchers and legal deposit libraries?

Research Context

The British Library began harvesting the UK web domain under legal deposit in 2013. The UK Web Archive had, by 2017, grown to 500Tb. However, UK legal deposit regulations, based on a centuries-old model of reading room access to deposited materials, affect the archive's significant potential for research: in practice, researchers can only access the full range of UK websites within the walls of selected institutions. DH scholars, though, require access to textual corpora and metadata in addition to interfaces for discovery and reading (Gooding, 2012). Winters argues that "it is the portability of data, its separability from an easy-to-use but necessarily limiting interface, which underpins much of the exciting work in the Digital Humanities" (2017: 246). Restricted deposit library access requires researchers to look elsewhere for portable web data: by undertaking their own web crawls, or by utilising datasets from *Common Crawl* (http://commoncrawl.org/) and the *Internet Archive* (https://archive.org). Both organisations provide vital services to researchers, and both innovate in areas that would traditionally fall under the deposit libraries' purview. They support their mission by exploring the boundaries of copyright, including exceptions for non-commercial text and data mining (Intellectual Property Office, 2014). This contrast between risk-enabled independent organisations and deposit libraries, described by interviewees as risk averse, challenges library/DH collaboration models such as *BL Labs* (http://labs.bl.uk) and Library of Congress Labs (https://labs.loc.gov).

Methodology

This paper analyses the impact of the UK regulatory environment upon DH reuse of the Legal Deposit UK Web Archive. It presents a quantitative analysis of information seeking behaviour, supported by insights from 30 interviews with UK legal deposit library practitioners. Quantitative datasets consisted of Google Analytics reports, and web logs of UK web archive usage, which were analysed in SPSS and Excel. These datasets allowed us to identify broad patterns of information-seeking behaviour.

Practitioner interviews were hand-coded to three levels in Nvivo: initial coding, to provide the foundations for higher level analysis; focused coding, to further refine the data; and axial coding, using the convergence of ideas as a basis for exploring the research questions (Hahn, 2008). This analysis will inform two further research phases: a broader quantitative analysis of UK ELD collections; and qualitative analysis of the ways that the research community, and DH researchers, use ELD collections.

Conclusion

This paper provides a vital case study of how legal deposit regulations can influence library/DH collaboration. It argues that UK ELD regulations use a print-era view of national collections to interpret digital preservation and access. A lack of media specificity, combined with a more cautious approach to text and data mining than allowed under UK copyright, restricts DH research: first, by limiting opportunities for innovative computational research; and second by excluding lab-based library/DH collaborative models. As web preservation activities become concentrated in a small group of key organisations, current regulations disadvantage libraries in comparison to not-for-profits, whose vital work is supported by an ability to take risks denied to legal deposit libraries. The UK's approach to national domain web archiving represents a lost opportunity for computational scholarship, requiring us to rethink legal deposit in light of the differing affordances of born-digital archives.

References

Black, M. L. (2016). The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. *International Journal of Humanities and Arts Computing*, 10(1): 95–109.

Brazier, C. (2016). Great Libraries? Good Libraries? Digital Collection Development and What it Means for Our Great Research Collections. In Baker, D. and Evans, W. (eds), *Digital Information Strategies: From Applications and Content to Libraries and People*. Waltham, MA: Chandos Publishing, pp. 41–56.

Brügger, N. (2012). Web History and the Web as a Historical Source. *Studies in Contemporary History*, 2 http://www.zeithistorische-forschungen.de/site/40209295/default.aspx (accessed 9 January 2017).

Gooding, P. (2012). Mass Digitization and the Garbage Dump: The Conflicting Needs of Quantitative and Qualitative Methods. *Literary and Linguistic Computing* doi:10.1093/llc/fqs054. http://llc.oxfordjournals.org/content/early/2012/12/22/llc.fqs054.abstract (accessed 30 July 2013).

Hahn, C. (2008). *Doing Qualitative Research Using Your Computer: A Practical Guide*. London: Sage Publications Ltd.

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria*, 25(1/2): 113–27.

Intellectual Property Office (2014). Exceptions to Copyright: Research *UK Government* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.
Winters, J. (2017). Coda: Web Archives for Humanities Research - Some Reflections. *The Web as History*. London: UCL Press, pp. 238–48.